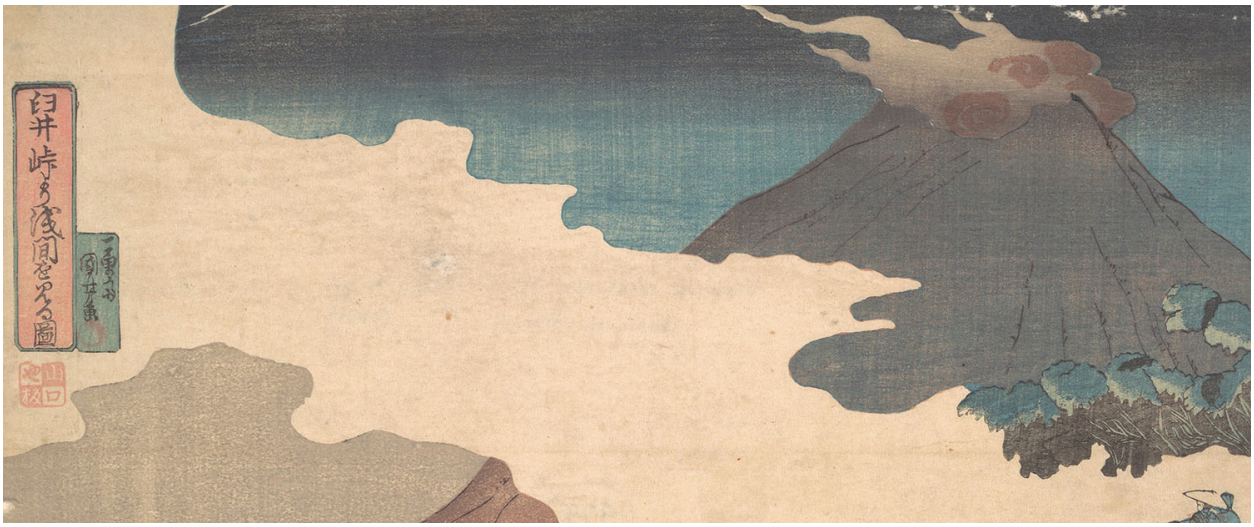


Hadoop HDFS Workbook

Learning Outcome

- Interact with the DFS shell by performing basic commands that create, append, read and delete files.

Pre-requisites

Dataset

- Download the dataset that we will use for learning Hadoop in the next lesson here https://files.grouplens.org/datasets/movielens/ml-100k.zip.
    - Unzip the zip file and find the u.data file containing ratings data with four columns (userID, MovieID, Rating, Timestamp). You can use the **u.item** file to find out what the records in u.data mean using the given metadata. You can explore these using a text editor.
- We will be using the IBM Analytics Engine on the IBM Cloud for interacting with Hadoop.
- The IBM Analytics Engine service provides us with ability to build and deploy Apache Hadoop clusters in minutes so you can focus on learning instead of wasting time setting up.

Sign Up for an IBM Cloud Account

The tasks that we will do will require the use of services on **IBM Cloud**.

We will be creating an IBM Cloud Trial account after which you will be able to access select IBM Cloud services and plans for 30-days at no charge, without the need to enter a credit card.

In case you are unable to complete the course in 30-days, and your trial has expired, you may need to create another trial account with a different email address.

Please follow the instructions below to create an account on IBM Cloud:

1. Go to https://cloud.ibm.com/registration/trial to create a trial account on IBM Cloud.
2. Enter your company Email address and a strong Password, as per criteria and then click the Next button.

   NOTE: Ensure that you provide an email address that you have not used previously to create any other IBM Cloud account, and you are able to readily access your email to retrieve the verification code required in the next step.
3. An email will be sent to the email address that you signed up with to confirm your email address. Check your email and copy and paste Verification code. Then click Next.
4. Once your email is successfully verified, enter your First name, Last name, and Country and click Next.
5. Go through the Account Notice and opt for Email updates if you desire, accept the terms and conditions and click on Continue.
6. Before creating your account, review the account privacy notice and acknowledge that you have read and understood by checking the checkbox and click on Continue.
7. You will be taken to the login screen. The username (which is your email address) is already populated.
8. Enter the password that you had chosen for the login. Unless you are using a private computer, don't choose to save the password.
9. Once you successfully login, you should see the dashboard.
10. Explore the IBM Catalog services and resources offered by IBM Cloud.

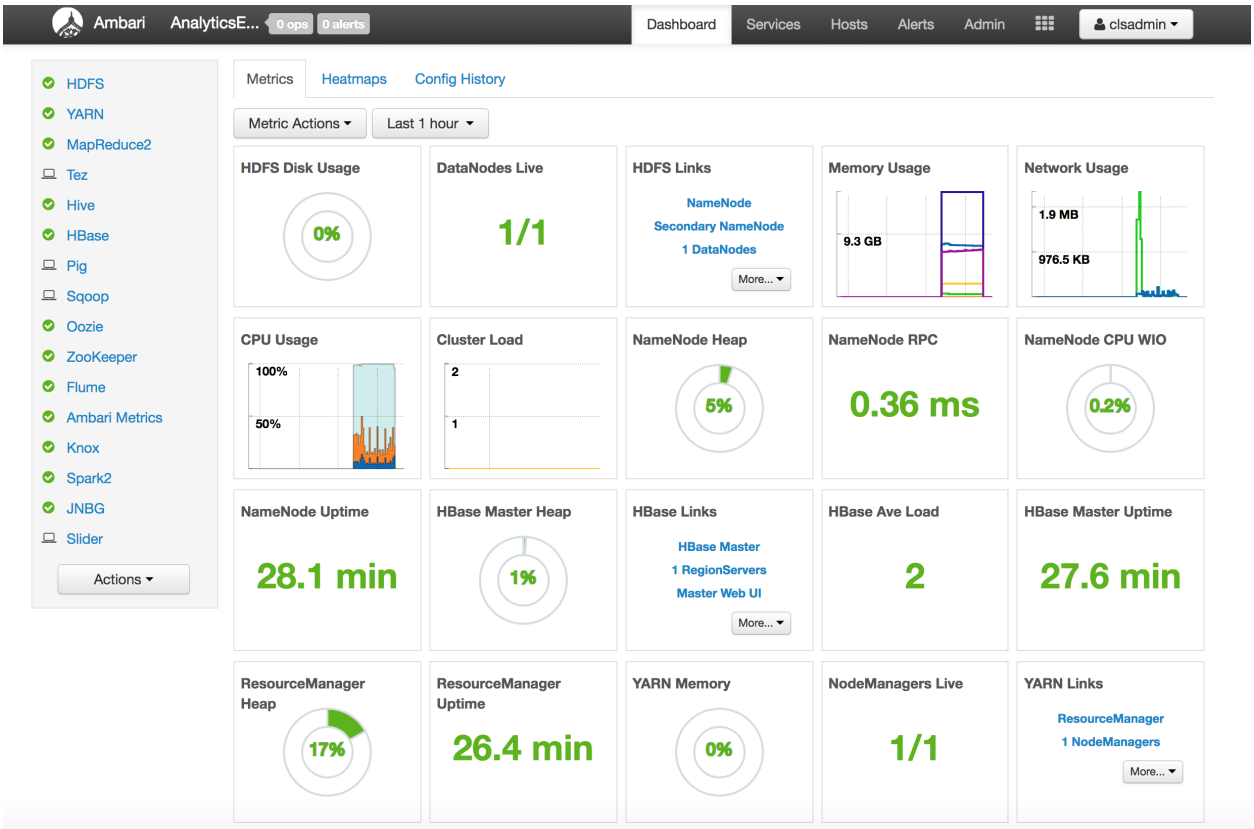
Deploying Hadoop on the Cloud

1. Use your account to log in IBM Cloud.
2. Click "**Catalog** and go to the page of IBM Cloud services.
3. Choose "Analytics" => "Analytics Engine", it goes to configuration of "Analytics Engine". Please specify the instance name, like 'Hadoop-Lab", and then:
   a. Select a region/location to deploy in.
   b. Select a resource group.
4. Then, click "Configure" and then set specify the parameters for your hadoop cluster:
   a. Select “Default” for Hardware Configuration
   b. Number of compute nodes: “1”
   c. Software package: “Spark and Hadoop”
5. Select “Create a service”

NOTICE: The provision process may take 30 ~ 50 minutes

6. Once the Analytics Engine service is created successfully, it's displayed in your dashboard like the below. Do be patient.
6. Manage Hadoop in Ambari by opening your Hadoop cluster from dashboard and then clicking "Launch Console" to log in Ambari console using the provided username and password.

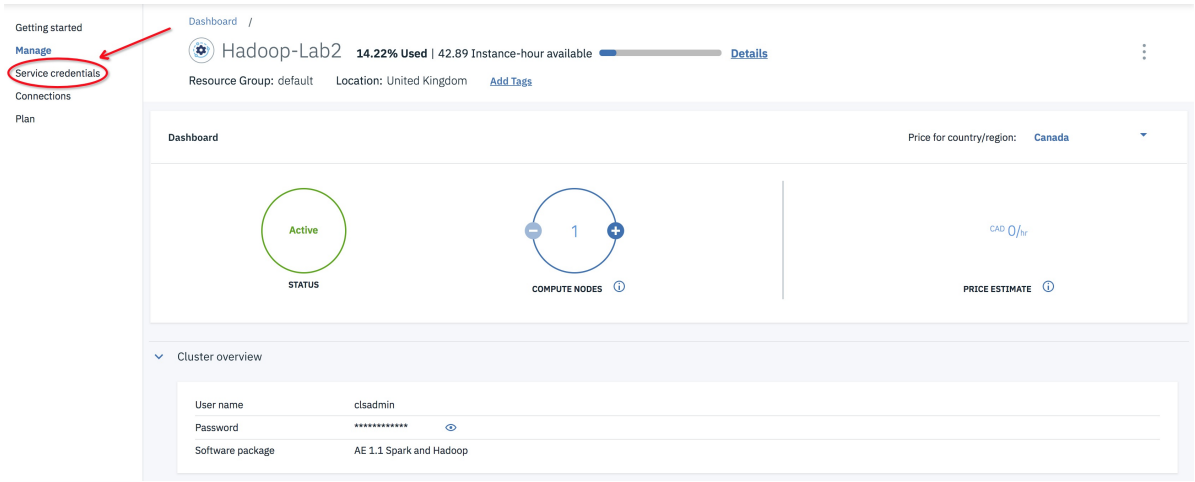
   NB: You might need to create a password for this on the Service credentials tab. The username to be used will be “clsadmin”.
7. After you log in Ambari console, the Dashboard of your Hadoop cluster is shown as below:



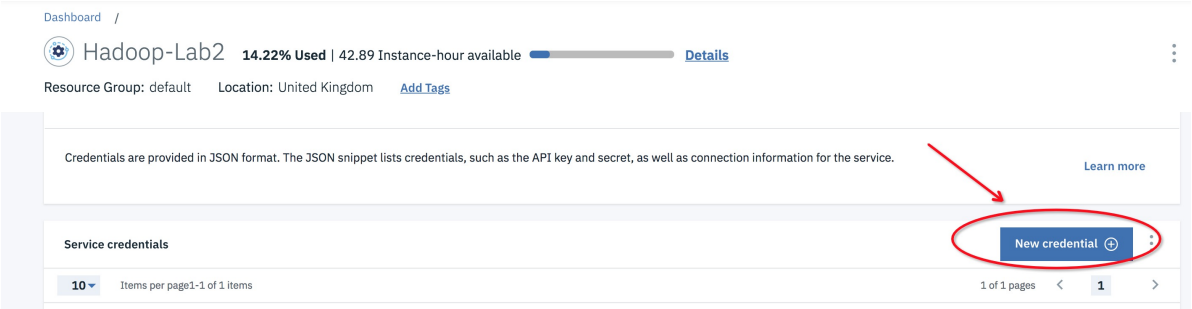
From the Dashboard, you are able to view all key metrics of your Hadoop cluster nodes and manage various services, like YARN, HDFS, Hive, etc.

9. To manage files on HDFS in the command line , you need to be able to log in the host from the remote ssh terminal . Here are steps on how to do it:

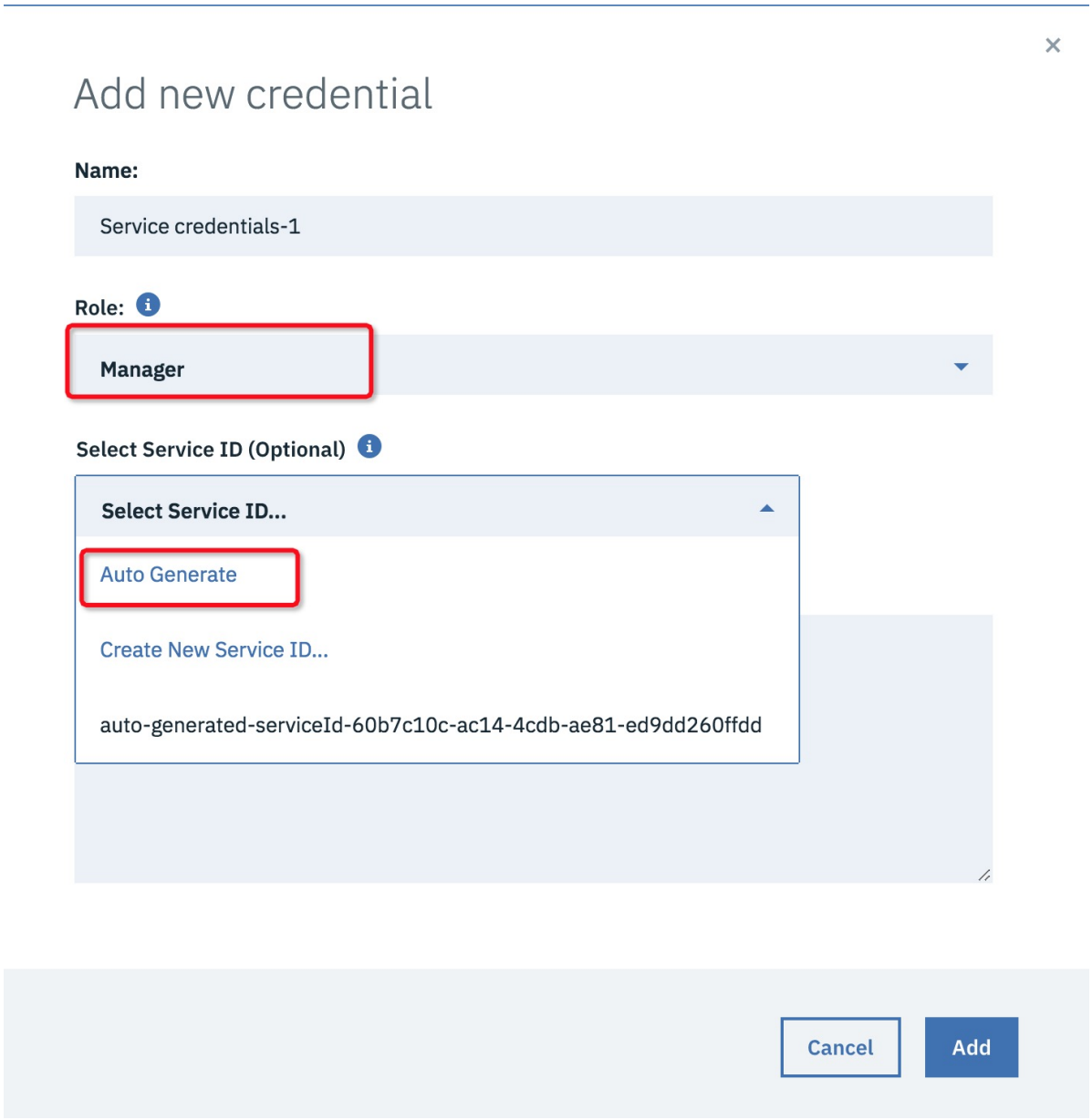
    a. Generate service credential for ssh by opening **Service credentials** on the left menu like the below:



b. Then, in **Service credentials** page , please choose **New credential** button to generate a credential:

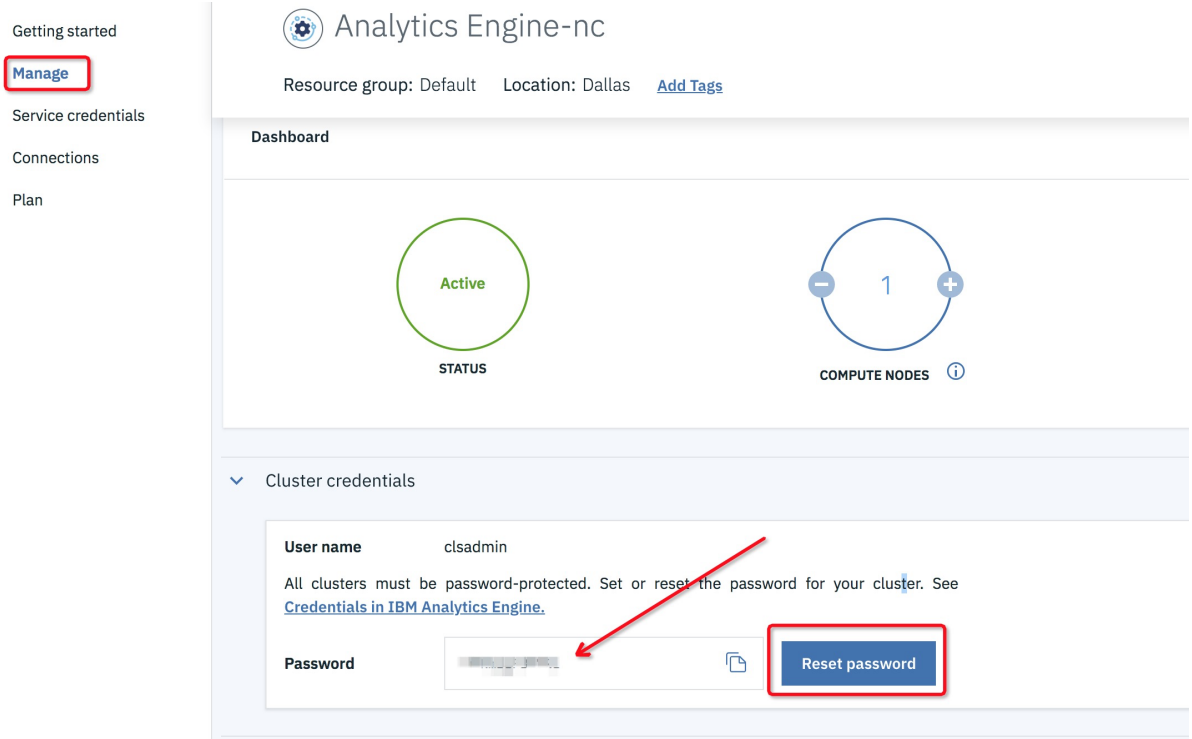


c. Select the values like the below to add a new credential:



Click '**Add**' button, a new credential is generated and then find the information for ssh remote login :

d. Then go to *Manage* and click *Reset password* to acquire a password:



10. Log in a cluster node via ssh:

From the above information, log in a cluster node via ssh client tool like a Terminal program, CMD Prompt, or PuTTY by providing username and password, like the below. We recommend using the default CMD Prompt available in your operating system.

```
leowu-macBookPro:train leowu$ ssh clsadmin@chs-jrn-551-mn003.eu-gb.ae.appdomain.cloud
The authenticity of host 'chs-jrn-551-mn003.eu-gb.ae.appdomain.cloud (169.50.228.221)' can't be established.
ECDSA key fingerprint is SHA256:LwCIvuaZQlNbkC9G+Nb573OyDXYm8gFMhImpnd3Suh4.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'chs-jrn-551-mn003.eu-gb.ae.appdomain.cloud,169.50.228.221' (ECDSA) to the list of known hosts.
clsadmin@chs-jrn-551-mn003.eu-gb.ae.appdomain.cloud's password:
```

Now, you are able to run HDFS commands like `**haddop fs -ls /**`:

```
[clsadmin@chs-jrn-551-mn003 ~]$ hadoop fs -ls /
Found 12 items
drwxrwxr-x   - ams      bihdfs          0 2018-10-04 16:46 /amshbase
drwxrwxrwx   - yarn     hadoop          0 2018-10-04 16:43 /app-logs
drwxr-xr-x   - hdfs     bihdfs          0 2018-10-04 16:45 /apps
drwxr-xr-x   - yarn     hadoop          0 2018-10-04 16:42 /ats
drwxr-xr-x   - hdfs     bihdfs          0 2018-10-04 16:43 /hdp
drwx------   - livy     bihdfs          0 2018-10-04 16:45 /livy2-recovery
drwxr-xr-x   - mapred   bihdfs          0 2018-10-04 16:43 /mapred
drwxrwxrwx   - mapred   hadoop          0 2018-10-04 16:43 /mr-history
drwx------   - clsadmin biusers         0 2018-10-04 16:53 /securedir
drwxrwxrwx   - spark    hadoop          0 2018-10-04 17:23 /spark2-history
drwxrwxrwx   - hdfs     bihdfs          0 2018-10-04 16:44 /tmp
drwxr-xr-x   - hdfs     bihdfs          0 2018-10-04 16:47 /user
[clsadmin@chs-jrn-551-mn003 ~]$
```

```
[clsadmin@chs-jrn-551-mn003 ~]$ hadoop fs -ls /hdp
Found 1 items
drwxr-xr-x   - hdfs bihdfs          0 2018-10-04 16:43 /hdp/apps
[clsadmin@chs-jrn-551-mn003 ~]$ hadoop fs -ls /hdp/apps
Found 1 items
drwxr-xr-x   - hdfs bihdfs          0 2018-10-04 16:45 /hdp/apps/2.6.5.0-292
[clsadmin@chs-jrn-551-mn003 ~]$
```

HDFS Commands

The **Hadoop Distributed File System (HDFS)** is the primary data storage system used by Hadoop applications.

HDFS is the primary or major component of the Hadoop ecosystem which is responsible for storing large data sets.

Hadoop dfs commands are the command-line utility for working with the Hadoop system.

These commands are widely used to process the data and related files. They are Linux based commands which control the Hadoop environment and data files.

Some of the commonly used Hadoop commands perform the following actions:

- List the directory structure to view the files and subdirectories.
- Create directory in the HDFS file system.

- Create empty files.
- Remove files and directories from HDFS.
- Copy files from other edge nodes to HDFS.
- Copy files from HDFS locations to edge nodes.

List of HDFS Basic Commands

To get familiar with basic HDFS commands, let's go through each of the following commands. Ensure that you run the commands so that you acquire hands on skills.

Help Command

`hdfs dfs -help`

Display

Displays a list of files and directories in the current path.

`hdfs dfs - ls`

List files in /user/hue directory.

`hdfs dfs -ls /user/hue`

Touch

Creates file will zero length in the current directory.

`hdfs dfs -touchz filename`

Cat

Displays the file content on the screen.

`hdfs -cat filename`

Make Directory

Create a new directory

`hdfs dfs -mkdir /newdirectory`

Copy files

Copy file or files

`hdfs dfs -cp filename filename2`

Move files

Move files or files from one location to another

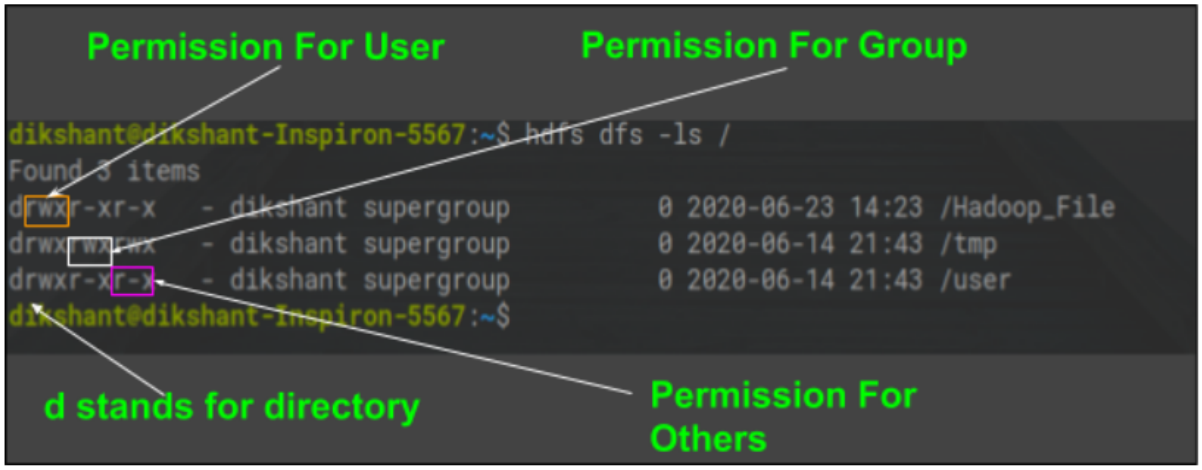
`hdfs dfs -mv filename /newdirectory/filename`

HDFS Basic Commands: Exercise

1. `hdfs dfs -help`
2. `hdfs dfs -help ls`
3. `hdfs dfs -ls`
4. `hdfs dfs -l /user`
5. `hdfs dfs -ls /user/hue`
6. `hdfs dfs -cat /user/hue/stocks.csv`
7. `hdfs dfs -cp /user/hue/stocks.csv /user/hue/stocks_copy.csv`
8. `clear`
9. `hdfs dfs -l /user/hue`
10. `hdfs dfs -cat /user/hue/stocks_copy.csv`
11. `hdfs dfs -mkdir /user/hue/hdfs-production`
12. `hdfs dfs -ls /user/hue`
13. `hdfs dfs -ls /user/hdfs-production`
14. `hdfs dfs -touchz /user/hue/hdfs-production/test.csv`
15. `hdfs dfs -ls /user/hue/hdfs-production`
16. `hdfs dfs -ls /user/hue/`
17. `hdfs dfs -mv /user/hue/stocks.csv /user/hue/hdfs-production/script.csv`
18. `hdfs dfs -ls /user/hue/hdfs-production`
19. `hdfs dfs -cat /user/hue/hdfs-production/test.csv`

HDFS Permission Commands

HDFS uses specific permissions model for file and directories. The user levels used in HDFS are Owner, Group and Others. For each of these users the following permissions are applicable: **read(r)**, **write(w)** and **execute(x)**.



List of HDFS Permission Commands

Change Mode

Change the mode file or files.

```
hdfs dfs -chmod [num] /somedirectory
hdfs dfs -chmod -R [num] /somedirectory
```

Change Group

Change the group association of file or files.

```
hdfs dfs -chgroup [group]/somedirectory
```

Change Owner

Change the ownership of file or files

```
hdfs dfs -chown [owner] /somedirectory
hdfs dfs -chown -R [owner] /somedirectory
```

HDFS Permission Commands: Exercise

1. `hdfs dfs -ls /user/hue/hdfs-production`
2. `hdfs dfs -chmod 700 /user/hue/hdfs-production/test.csv`
3. `hdfs dfs -ls /user/hue/hdfs-production`
4. `hdfs dfs -mkdir /usr/hue/hdfs-production/inside`
5. `hdfs dfs -ls /user/hue/hdfs-production`
6. `hdfs dfs -chmod -R 700 /user/hue/hdfs-production/inside`

7. `hdfs dfs -ls /user/hue/hdfs-production`
8. `hdfs dfs -mv /user/hue/hdfs-production/test.csv`
9. `hdfs dfs -ls /user/hue/hdfs-production/inside`
10. `hdfs dfs -chgrp -R student /user/hue/hdfs-production/inside`
11. `hdfs dfs -ls /user/hue/hdfs-production`
12. `hdfs -dfs -l /user/hue/hdfs-production/inside`
13. `clear`
14. `hdfs dfs -chmod -R 644 /user/hue/hdfs-production/inside`
15. `hdfs dfs -ls /user/hue/hdfs-production/inside`
16. `hdfs dfs -chown -R hue /user/hue/hdfs-production/inside`

Moving Data Commands

List of Moving Data Commands

Put Command

Copy a file or files form local system into hdfs.

`hdfs dfs -put /tmp/filename /user/hdfs/filename`

Get Command

Copy file or file sfrom HDFS to local file system.

`hdfs dfs -get /user/hdfs/filename /tmp/filename`

Move file from Local

Moves files from local directory into a cluster and deletes local copy.

`hdfs dfs -moveFromLocal /tmp/filename /user/hdfs/filename`

Moving Data Commands: Exercise

1. `ls /tmp`
2. `hfds dfs -ls /user/`
3. `hdfs dfs -ls /user/ student`

4. `hdfs dfs -ls /user/student`
5. `hdfs dfs -put /tmp/stocks.csv /user/student/stocks.csv`
6. `hdfs dfs -ls /user/student`
7. `clear`
8. `hdfs dfs -movefromlocal /tmp/stocks.csv /user/student/stocks_test.csv`
9. `ls /tmp`
10. `hdfs dfs -ls /user/student`
11. `hdfs dfs -get /user/student/stocks.csv /tmp/stock.csv`
12. `ls /tmp`
13. `hdfs dfs -ls /user/student`

Maintenance Commands

List of Maintenance Commands

Remove a Directory

Moves a directory into trash

`hdfs dfs -rm -r directory`

Expunge Command

Empties trash

`hdfs dfs -expunge`

HDFS maintenance Commands Exercise

1. `hdfs dfs -ls /user/student`
2. `hdfs dfs -rm /user/student/stocks.csv`
3. `hdfs dfs -ls /user/student`
4. `hdfs dfs -rm r /user/hue/hdfs-production`
5. `clear`

6. `hdfs dfs -ls`
7. `hdfs dfs -ls .Trash`
8. `hdfs dfs -ls .Trash/Current/user`
9. `hdsf dfs -expunge`
10. `hdfs dfs -ls .Trash/Current/user`